# Helping Students Know 'Further' – Increasing the Flexibility of Students' Knowledge Using Symbolic Invention Tasks

**Ido Roll (idoroll@cmu.edu)**
Human Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

**Vincent Aleven (aleven@cs.cmu.edu)**
Human Computer Interaction Institute, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

**Kenneth R. Koedinger (koedinger@cmu.edu)**
Human Computer Interaction Institute and Department of Psychology, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

**Abstract**

Invention as Preparation for Learning (IPL) is a teaching strategy in which students attempt to develop novel solutions prior to receiving instruction (Schwartz & Taylor, 2004). This method was previously shown to prepare students to learn independently from future learning opportunities that build upon the materials learned in class. We began unpacking the IPL process by identifying its components and evaluating the contribution of *generative reasoning* (in the form of symbolic invention) on top of *comparative reasoning* (in the form of ranking alternatives). An in-vivo study in 6 middle-school classes with 105 students found that generative reasoning is an essential component of IPL. Furthermore, we found that students who attempted to invent symbolic models during the IPL process (generative reasoning) were able to invent new strategies during the post-test. At the same time, students who completed the IPL process without designing symbolic methods were in need for worked-out examples in order to solve new-strategy problems in the post-test. We propose a mechanism that explains how invention leads to the observed increased flexibility in students' knowledge.

**Keywords:** Invention as Preparation for Learning; Preparation for Future learning; Transfer; Comfort Zone; Generative Reasoning.

## Introduction

Invention as Preparation for Learning (IPL) is a teaching strategy that uses constructivist instructional methods and direct instruction in a complementary fashion (Schwartz & Martin, 2004). First, students are asked to invent general methods (and their mathematical expressions) to evaluate a set of examples with regard to one aspect of the data. Figure 1 shows an example of such a task, in which students are asked to invent a method for comparing the variability of two datasets, in order to choose the more consistent one (i.e., where data is "closer together"). Following the invention attempt, students receive direct instruction on canonical methods and practice them. For example, following the task detailed in Figure 1, students receive instruction on Mean Absolute Deviation and practice applying it. While students often fail to invent general valid methods, research suggests that this experience prepares them to better learn independently from learning opportunities that follow the instruction (Schwartz & Martin, 2004; Kapur, 2008).

> The Bouncers Trampoline Company tests their trampolines by dropping a 100 lb weight from 15 feet. They measure how many feet the weight bounces back into the air. They do several trials for each trampoline. Here are the results for two of their trampolines:
>
> Trampoline A: {1 3 5 7 9}
> Trampoline B: {3 4 5 6 7}
>
> Which trampoline is more consistent, that is, its test results are closer together?

Figure 1: The trampoline IPL task.

IPL tasks use contrasting cases to direct students' attention to deep features of the domain. Rather than analyzing a single dataset, as commonly done in show-and-practice problems, IPL tasks ask students to compare two or more sets of data that vary along a single deep feature. For example, the two sets in Figure 1 have the same average and sample size but differ in their range. The contrasting cases also give students a baseline against which to evaluate their inventions, since their intuitive comparison of the cases is often clear and correct (Schwartz, Sears, & Chang, 2007). The invention activity itself, prior to instruction, can be divided into two parts. First, students analyze the contrasting cases and rank them intuitively according to the target construct (e.g., variability). We refer to this stage as *comparative reasoning*, since students reason about the task by comparing the different cases. The second part is the design of mathematical methods, in which students attempt to invent general valid methods that rank the cases in the same way as their intuitive ranking. We refer to this stage as *generative reasoning*, since students generate symbolic methods to quantitatively compare the contrasting cases. Table 1 shows a summary of the IPL process.

In the current study, we begin to unpack the IPL process and its effects. Our first research question evaluates the different roles of comparative vs. generative

Table 1: The IPL process and experimental conditions.

| Activity type: | Example task: | Experimental conditions: Full IPL | No Design |
|---|---|---|---|
| **Invention:** | | | |
| Comparative reasoning | “Rank the following trampolines according to their consistency” | ✓ | ✓ |
| Generative reasoning | "Invent a general mathematical method that yields a similar ranking" | ✓ | |
| **Show and practice:** | | | |
| Direct Instruction | “One method that mathematicians use is Mean Absolute Deviation…” | ✓ | ✓ |
| Practice | "Apply the canonical method to the following problems:" | ✓ | ✓ |

reasoning in the IPL process. Will students who are engaged in both types of reasoning (that is, ranking followed by design) show superior learning compared to students who are engaged in comparative reasoning alone (that is, ranking only) prior to instruction?

One hypothesis argues that generative reasoning (in the form of symbolic invention) is necessary to improve encoding of subsequent instruction. First, generative reasoning facilitates a process in which students express their prior ideas, identify their shortcomings, and refine their mental models, thus enabling conceptual change (Smith, diSessa, & Roschelle, 1994). For example, the self-explanation literature shows that asking students to explain their errors facilitates conceptual shift (c.f., Siegler, 2002).

By attempting to invent and understand how different symbolic procedures succeed (or fail) to capture the differences between the contrasting cases, students also acquire a more cohesive and integrated understanding of the deep features of the domain. The importance of the symbolic nature of the process was demonstrated by Schwartz, Martin, and Pfaffman (2005), who asked students to reason verbally or mathematically about the balance beam problem. All students noticed the deep features of the balance beam domain - distance and weight. However, only students who reasoned mathematically were able to reconcile the two dimensions to a single representation. Interestingly, students’ thinking evolved even though their solutions were not complete, similar to the IPL effect.

Lastly, the generative reasoning process may help students understand the function of the different components of the procedure (for example, dividing by N controls for sample size). Thus, students may encode the subsequent instruction by function and not merely by procedure. Functional mental models were previously shown to lead to better adaptation of knowledge (Kieras & Bovair, 1984). Hatano and Inagaki (1986) describe a similar process in which developing mental models of how procedures interact with empirical knowledge helps students acquire conceptual understanding of the domain.

An alternative hypothesis argues that comparative reasoning is sufficient to achieve the learning benefits of IPL. According to this hypothesis, the benefits of invention stem from noticing and encoding the deep features of the domain. The comparative reasoning activity achieves that benefit by asking students to compare contrasting cases that differ with respect to their deep features. (Bransford & Schwartz, 2001). This qualitative analysis helps students set requirements for a valid model and thus acquire a better understanding (even if implicit) of the target concepts. Furthermore, according to this hypothesis, not only does the symbolic invention *not* contribute to future learning, it may waste students’ time (and thus reduce efficiency) or impose excessive cognitive load (Kirschner, Sweller & Clark, 2006).

A second research question addressed by our current study examines the effect of IPL on the flexibility of students’ knowledge. We follow a distinction made by McDaniel and Schlager (1990) between transfer problems that require the application of a learned strategy (conventional transfer problems) and transfer problems that require the generation of a new strategy. McDaniel and Schlager found that while discovery tasks improve students’ performance on the latter, they have no effect on conventional transfer problems. Schwartz and Martin (2004) add a twist to these results. They found that IPL improves students’ ability to solve new-strategy problems as long as they are provided with instruction on how to do so. To further investigate the effect of IPL on knowledge flexibility, we evaluate students’ ability to independently solve new-strategy problems and encode new-strategy instructions. Our hypothesis, as supported by McDaniel and Schlager (1990), is that students who are engaged in IPL will acquire more flexible knowledge and thus will demonstrate better performance on new-strategy items. At the same time they will not show better ability to use existing strategies in novel contexts (conventional transfer items). Furthermore, following the findings of Schwartz and Martin (2004), we hypothesize that the effect of IPL will be mainly on encoding new-strategy instructions.

# Methods

## Design

The study compared two conditions, as seen in Table 1: *Full IPL* and *No Design*. Students in both conditions received contrasting cases and were asked to rank them according to the target concept (comparative reasoning). This phase was followed by a class discussion of the correct ranking. All students also received direct instruction (procedural and conceptual) and opportunities for practice. The two conditions differed with regard to the invention activity:

*Full IPL students* were asked to design mathematical methods for ranking the cases (generative reasoning).

This design activity followed the discussion of the contrasting cases and came before the direct instruction. The Full IPL condition resembled the instruction tested by Schwartz and Martin (2004). The design process had two distinct iterative stages: First, students invented general mathematical procedures or visual representations that, when applied to the cases, should yield rankings similar to their (intuitive) predictions. Then, students evaluated their methods by comparing the rankings generated by their designed methods to their predictions. When their methods produced the desired ranking, students moved on to the next set of contrasting cases (each problem included several sets of contrasting cases, emphasizing different features of the domain, such as range, number of points, central tendency vs. distribution, etc). A mismatch in the ranking led to an iterative debugging process, in which students attempted to identify the reason for the failure of their model and improve it. This process was chosen for several reasons. First, the process seems to match students' natural approach to the IPL task, as evaluated during our pilot studies. Second, these steps match the hypothetico-deductive scientific method, and thus help students practice an important set of skills (Popper, 1963). Third, the scientific method was shown to transfer well across domains and tasks (Rivers & Vockell, 1987), especially when applied iteratively to debugging procedures using evidence (Carver, 1998).

*No Design students* received instruction immediately following the ranking and the class discussion. Instead of a design stage, they received more comprehensive instruction and practice. The contrasting cases were used during the instruction to demonstrate the canonical procedure, and students evaluated the canonical procedure against their predictions. The No Design condition resembled traditional direct instruction with the addition of a short, guided comparative reasoning activity using contrasting cases.

## Participants

The study took place in six 7th grade classes at a public middle school in the Pittsburgh area (30% free lunch, 35% minorities). Three of the classes were regular classes and three were advanced (pre-Algebra classes). Since the activities varied significantly between conditions, we could not assign students to conditions within class. Instead, we assigned whole classes to conditions. In both levels, two classes were randomly assigned to the IPL condition and one to the No Design condition. In order to minimize the chances for selection bias we validated that the end-of-year and standardized-tests scores did not differ between classes. The study included two topics. Due to absentees, not all students participated in both topics. 96 students participated in the first topic (66 in Full IPL, 30 in No Design, split rather evenly between regular and advanced classes). 78 students participated in the second topic of the study (45 in Full IPL, 33 in No Design). Notably, more than half of the advanced students in the Full IPL condition missed the second topic due to an overlapping activity.

## Materials

The study included two topics: (1) central tendency and graphing (histograms, stem and leaf plots, bar charts, box and whisker, mean, median, mode and range) and (2) variability (distribution, consistency, mean absolute deviation).

Each of the topics included two problems with multiple sets of contrasting cases. The two problems for central tendency and graphing asked students to choose which class to attend (based on test scores) and which gender shops more (based on revenue data). The two problems for variability asked students to identify which trampoline is more consistent (based on factory testing data) and which rocket is more predictable (based on NASA tests). The contrasting cases were identical in both conditions. All students encountered them in the comparative reasoning phase and the instruction phase. In addition, the Full IPL students used them as basis for invention. All materials were piloted in the lab and in another class from the same cohort in the school.

To evaluate the effect of condition on students' knowledge flexibility we used several types of transfer items (in addition to normal items; see Table 2). The first type, conventional transfer items, required the application of knowledge taught in class in a new context. For example, students learned in class how to use histograms and stem-and-leaf plots. One conventional transfer item asked students to match between different representations of the same data without explicitly going through the data table. While this was a new type of problem, the skills learned in class were sufficient for its solution.

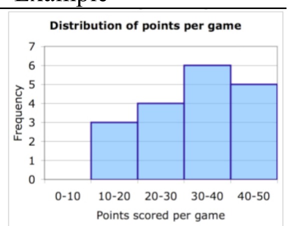
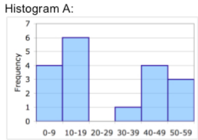
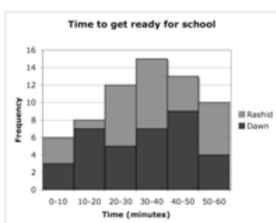
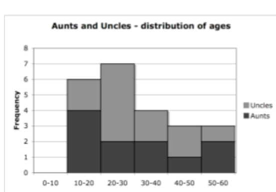
The second type of transfer items required the generation of a new strategy during the test. These strategies built upon, but extended beyond, the materials learned in class. For example, students in the class learned how to interpret conventional histograms that represent a single set of data. A new-strategy item asked students to interpret histograms with two stacked sets of data.

Each test form in the graphing post-test included two new-strategy items. One of the items had no additional instruction, and thus evaluated students' ability to adapt their knowledge spontaneously. The other item (counter balanced between forms) followed an embedded learning resource in the form of solved examples with comprehension questions. These items, termed future learning items (Bransford & Schwartz 2001), evaluated students' ability to comprehend additional instruction and apply it to new-strategy problems without further assistance. There were at least 3 items in between each learning resource (solved example) and the corresponding future learning item. The combination of the two types of new-strategy items (with or without learning resource) allows us to evaluate two aspects of knowledge flexibility: the ability to encode and apply new instruction, and the ability to spontaneously generate the relevant strategy without additional instruction. The tests included also motivational and metacognitive assessments. However, these are outside the scope of the current paper.

### Procedure

The study spanned 4 days with two periods per day. The first two days covered topics of central tendency and graphing. The subsequent two days were on variability. Both topics followed a similar structure. Full IPL students completed the invention activities on days 1 and 3, and received instruction and practice on days 2 and 4. No Design students received instruction and practice on all four days. On day 1, all students completed a pre-test on central tendency and graphing (no pre-test on variability was given under the assumption of a floor effect). Post-tests on each topic were administered at the end of the relevant practice on day 2 (graphing posttest) and day 4 (variability posttest). Students completed a delayed post-test about a month after the study.

Table 2: Types of assessments used in the study.

| Item type | Example |
|---|---|
| **Normal test items:**<br>Test items on topics covered during instruction<br><br>Example: *In how many games did the team score between 20 and 30 points?* |  |
| **Conventional transfer items:**<br>Items that require to apply existing knowledge in new context<br><br>Example: *True or false: The stem and leaf plot and Histogram A show the same data* | Stem \| Leaf<br>0 \| 3 8<br>1 \| 5 8 9<br>2 \| 1 2 4 4 6<br>3 \| 2 6 7 8<br>4 \| 2 5<br>5 \| 7<br><br>Histogram A:<br> |
| **New strategy items:**<br>Items that require a new strategy, different from what was learned in class. For example, students did not learn how to read histograms with two datasets and thus needed to make sense of it by themselves.<br><br>Example: *How many of Dawn's friends take less than 10 minutes to get ready for school?* |  |
| **Embedded instruction:**<br>Half of the new-strategy items followed a solved example embedded in the test. The solved example illustrated the relevant new strategy.<br><br>Example: *How many aunts are between 30 and 40 years old? Answer: 2 aunts. We look only at the darker gray that represents aunts.* |  |

## Results

There were no significant differences between groups on pre-test (Full IPL=33%, No Design=36%, $F(4,97)=9.7$, $p<.2$). A repeated-measures analysis on identical items between the pre- and post-tests showed significant learning ($F(4,87)=120.6$, $p<.0005$). Figure 2 summarizes the results of the different measures.

### Normal measures

An ANCOVA of students' performance on normal items on the graphing post-test (controlling for performance at pre-test) found no main effect for condition, but a significant interaction between condition and class-level ($F(4,90)=22$, $p<.05$). A separate ANCOVA for each class level showed that in the regular classes Full IPL students did marginally significantly better than No Design students (50% vs. 43% respectively, $F(2,38)=2.9$, $p<.1$). There was no difference between conditions in the advanced classes.

A similar analysis in the variability post-test showed a marginally significant interaction between condition and class-level ($F(4,73)=3.4$, $p<.07$). Analysis within the levels found that No Design students did marginally significantly better in the regular classes (69% vs. 48%, $F(2,37)=2.9$, $p<.1$). There were no significant differences between conditions in the advanced classes.

Students in both conditions did equally well on conventional transfer items in both topics.

### New strategy measures

The graphing post-test included new-strategy items with and without embedded learning resources. An ANCOVA of students' performance on new-strategy items without learning resources (controlling for performance at pre-test) found a significant advantage for Full IPL students ($F(90)=5.3$, $p<.03$). There is also significant interaction between condition and class-level on these items ($F(4,90)=3.8$, $p=.05$). A separate ANCOVA for each class level reveals a significant effect only for advanced students ($F(2,51)=7.9$, $p<.01$). Notably, the effect holds also when controlling for performance on normal items on the same post-test ($F(2,51)=6.4$, $p=.01$). Furthermore, while No Design students showed a significant drop in performance on no-strategy items in the absence of instruction ($t(15)=2.4$, $p<.03$), the scores of Full IPL students on future learning

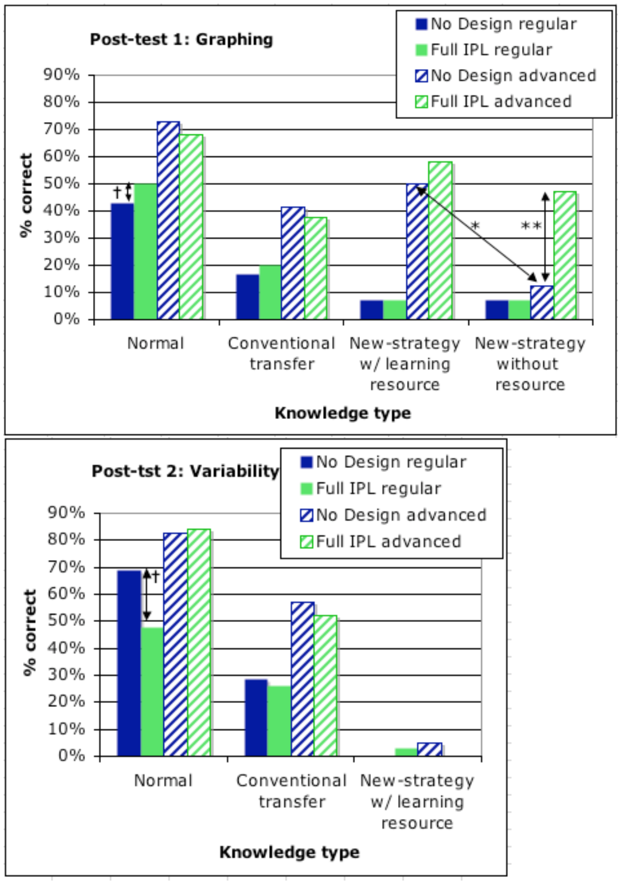



Figure 2: Performance on post-tests as a function of class-level and condition († - $p<.1$; * - $p<.05$; ** - $p<.01$)

items were not affected significantly by removing the learning resources ($t(37)=1.0$, $p>.3$).

Due to a somewhat unfortunate decision, the variability post-test included only new-strategy items that followed embedded learning resources. Scores on these items were at floor (2% for Full IPL students, 3% for No Design students). There was no significant effect for condition or its interactions on performance on these items.

## Discussion

Regarding our first research question, we found that generative reasoning (on top of comparative reasoning) had a positive effect on students' ability to solve new-strategy problems with no learning resource in the advanced classes. At the same time, as hypothesized, it had a marginal effect on normal or conventional transfer items. These results are interesting especially since Full IPL students had approximately half the time for instruction and practice compared with their No Design counterparts.

Regarding the second research question, which dealt with students' knowledge flexibility, we found that in the advanced classes, students who designed novel methods during IPL were more capable of solving problems that require the use of novel strategies. This finding echoes the effect found by McDaniel and Schlager (1990). Interestingly, the effect of IPL on new-strategy items with no resources holds even when controlling for performance on normal items on the same test. Thus, this effect can probably *not* be attributed to more domain knowledge. Instead, it is likely the outcome of a different encoding of domain knowledge, in a manner that is not reflected in normal or transfer items.

On further scrutiny, students in both conditions did equally well on all tasks for which they received some form of instruction - whether in class (on normal and conventional transfer items) or embedded in the test (on new-strategy items with embedded learning resources). Regarding the latter, it seems that Full IPL students did not need the additional instruction whereas No Design students did not manage to solve the new-strategy problems without it. The performance of Full IPL students on new-strategy items remained virtually the same even in the absence of embedded instruction. This finding is at odds with earlier findings by Schwartz and Taylor (2004) who found that IPL improves students' ability to encode future instruction but not solve novel problems without additional instruction. One explanation for the discrepancy between the studies is that the control group in Schwartz and Taylor (2004) did not engage in comparative reasoning. Therefore, it may be that the comparative reasoning stage helped students in our study to encode the novel instruction.

An alternative explanation examines these results in terms of 'distance' from original classroom instruction. It may be that the embedded instruction on the first topic in our study was close to the classroom material, and thus simple enough for all students to encode. In contrast, the embedded learning resource in the study described by Schwartz and Martin (2004) was sufficiently far from the classroom instruction. Therefore, only IPL students, who had acquired more flexible knowledge, could learn from it and apply the acquired knowledge successfully. This explanation further suggests that in the absence of additional instruction, only Full IPL students in our study could make the leap and answer the target new-strategy items.

While this argument explains performance on new-strategy items (with or without instruction) in terms of distance from classroom instruction, it does not explain what factors determine this distance. What makes some items 'closer' than others? What prepared Full IPL students for improved performance on some items but not on others?

Students may grapple with many challenges during the invention phase, many of which do not receive attention during classroom instruction. Students who invent are exposed to various challenges by virtue of attempting to invent general valid methods. We hypothesize that students use knowledge acquired during these experiences when later integrating new-strategy tasks into their existing body of knowledge. For example, the post-tests in this study included three new-strategy items, requiring the following new strategies: (1) comparing multiple datasets in a single representation; (2) representing data in unconventional intervals; and (3) finding the ratio between variability and average in order to account for differences in magnitude. These topics were not covered during classroom instruction. However, when we analyzed students' inventions, we noticed that many inventions included features that could prepare students to expand the instructed knowledge and invent the first two strategies (see Figure 3). Subsequently, Full IPL students demonstrated better performance on the relevant new-strategy items. At the same time, no student attempted during invention to compare datasets with different magnitudes. Correspondingly, Full IPL students did not exhibit better performance on this new-strategy item.

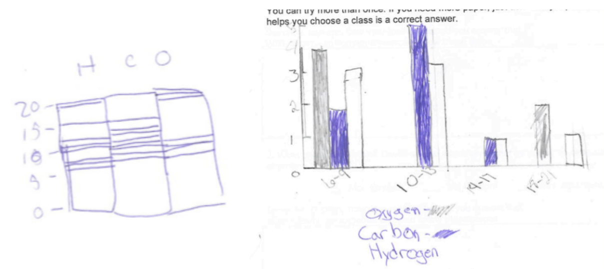



Figure 3: Inventions by students comparing classes based on test scores. While not mathematically valid, such inventions may prepare students to spontaneously develop new strategies such as comparing components of data and splitting data to bins other than by 10's.

In summary, we identified two components of the IPL process: comparative reasoning and generative reasoning. We found that generative reasoning (in the form of symbolic inventions) led to more flexible knowledge and thus is an essential component of the IPL process. Our results further show that students who invent during IPL are more likely to invent successfully during subsequent tests. Notably, these benefits for IPL were found even though none of the students invented a mathematically sound method during the invention phase. In addition, this effect holds even when controlling for domain knowledge (as assessed by normal items). More studies are needed to better understand the form of the knowledge acquired during IPL and to predict a-priori on which tasks IPL instruction shows benefits.

## Acknowledgments

We thank Dan Schwartz, David Klahr, Ofira Roll, Jo Bodnar, Gail Kusbit, Beth McCalister, Tom Kendro, Ed Wellman, Hampton Conway, and Audrey Russo for their help in conducting this research. This work is supported by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420.

## References

Bransford, J. D., & Schwartz, D. L. (2001). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24(3), 61-100.

Carver, S. M. (1998). Learning and transfer of debugging skills: Applying task analysis to curriculum design and assessment. In R. E. Mayer (Ed.), *Teaching and learning computer programming: Multiple research perspectives.* (pp. 259-97). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan*. (pp. 262-72). NY: Freeman.

Kapur, M. (2008). Productive failure. *Cognition and Instruction,* 26(3), 379 - 424.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, 8(3), 255-73.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75-86.

McDaniel, M. A., & Schlager, M. S. (1990). Discovery learning and transfer of problem-solving skills. *Cognition and Instruction*, 7(2), 129-159.

Popper, K. R. (2002). *Conjectures and refutations : The growth of scientific knowledge* . London ; New York : Routledge,. (Original work published 1963)

Rivers, R. H., & Vockell, E. (1987). Computer simulations to stimulate scientific problem solving. *Journal of Research in Science Teaching,* 24(5), 403-415.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22(2), 129-184.

Schwartz, D. L., Martin, T., & Pfaffman, J. (2005). How mathematics propels the development of physical knowledge. *Journal of Cognition and Development*, 6(1), 65-88.

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. In M. C. Lovett, & P. Shah (Eds.), *Thinking with data*. (pp. 319-44). New York : Routledge.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott, & J. Parziale (Eds.), *Microdevelopment - transition processes in development and learning*. Cambridge University Press.

Smith III, J. P., diSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, 3(2), 115-163.